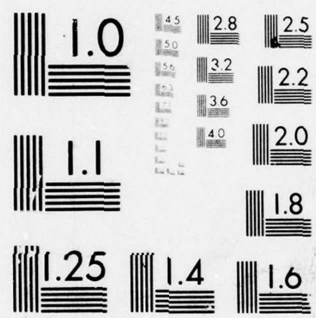
AD-A064 739 AIR FORCE HUMAN RESOURCES LAB BROOKS AFB TEX F/G 5/10
ESTIMATING ITEM CHARACTERISTICS CURVES.(U)
NOV 78 M J REE
UNCLASSIFIED AFHRL-TR-78-68 NL

1 OF 1
AD
A0 64739



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFHRL-TR-78-68

LEVEL

AIR FORCE

HUMAN RESOURCES

ESTIMATING ITEM CHARACTERISTIC CURVES

By

Malcolm James Ree

PERSONNEL RESEARCH DIVISION
Brooks Air Force Base, Texas 78235

November 1978
Final Report for Period March 1978 – September 1978

LABORATORY

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Personnel Research Division, under project 7719, with HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235. Dr. Malcolm James Ree (PES) was the Laboratory Principal Investigator.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved for publication.

LELAND D. BROKAW, Technical Director
Personnel Research Division

RONALD W. TERRY, Colonel, USAF
Commander

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFHRL-TR-78-68	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): ESTIMATING ITEM CHARACTERISTIC CURVES	5. TYPE OF REPORT & PERIOD COVERED: Final March 1978 – September 1978
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Malcolm James Ree	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Personnel Research Division, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas 78235	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62703F 77191515
11. CONTROLLING OFFICE NAME AND ADDRESS: HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235	12. REPORT DATE: November 1978
	13. NUMBER OF PAGES: 18
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

item analysis
item characteristic curves
latent trait theory
maximum likelihood
psychometrics
simulations
test construction

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A simulation study of the effectiveness of the following four item characteristic curve estimation programs was conducted: ANCILLES, OGIVIA (from U. S. Civil Service Commission); LOGIST (from Educational Testing Service); and simple transformations to the item-test biserial correlation. Using the three-parameter logistic model, three groups of 2,000 simulated subjects were administered 80-item tests. These simulated item responses were then calibrated using the four programs. The estimated item parameters were compared to the known item parameters in four analyses for each program in all of the three data sets. It was concluded that the selection of an item calibration procedure should be dependent on the distribution of ability in the calibration sample, the planned uses of the item prameters, and the computer resources available.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

404 425

LB

SECURITY CLASSIFICATION OF THIS PAGE(*When Data Entered*)

SECURITY CLASSIFICATION OF THIS PAGE(*When Data Entered*)

PREFACE

This research was conducted under project 7719, Air Force Development of Selection, Assignment, Performance Evaluation, Retention and Utilization Devices; task 771915, Perceptual and Computer managed measurement.

The author wishes to extend his appreciation to Dr. Vern Urry, United States Civil Service Commission, and to Dr. Frederick M. Lord, Educational Testing Service, for making their computer programs available and for their suggestions concerning the simulation and analyses. James R. McBride, Naval Personnel Research and Development Center, James B. Sympson, University of Minnesota, and Vincent Maurelli, Army Research Institute, provided much appreciated assistance in the conduct of this study.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Page 4 in this document is left out intentionally

ESTIMATING ITEM CHARACTERISTIC CURVES

I. INTRODUCTION

Increased interest in computer-driven adaptive testing, automated item banking, and automated test construction has made the estimation of the Item Characteristic Curve (ICC) important. This curve describes the relationship between the ability of individuals and the probability of their answering a test question correctly. It is useful in estimating test scores, equating the scores of various tests, and scoring responses during adaptive testing. There are several methods for estimating ICC within available computer programs. Selection and implementation of the appropriate program becomes a task for the practitioner. The objective of this study is to compare the merits of four available computer programs.

The Research Problem

In order to estimate an ICC, a conceptual model must be defined and item parameters must be estimated. The three-parameter logistic model of Birnbaum (Lord & Novick, 1968) is the most frequently used for relating item responses to subjects' ability. The three parameters, a, b, and c, are item discrimination, item difficulty (or location), and probability of chance success (or lower asymptote), respectively.

The curve described by these parameters takes the shape of an (cumulative frequency) ogive or an "s" with the upper asymptote approaching a probability of 1.0 and usually a lower asymptote of a probability greater than 0.0. The ogive describes the probability of obtaining a correct answer to an item as a monotonic increasing function of ability.

The item discrimination parameter (a) is a function of the slope of the ICC and generally ranges from .5 to about 2.5. The value of a equal to about 1.0 is typical of many test items, while a values below .5 are insufficiently discriminating for most testing purposes, and a values above 2.0 are infrequently found.

The item difficulty parameter (b) describes the point of inflection of the ICC and is usually scaled between −2.5 and +2.5 although the metric is arbitrary.

The item guessing parameter (c) is the lower asymptote of the ICC and is generally conceived to be the probability of selecting the correct item-option by chance alone. Most test items have c parameters greater than 0.0 and less than or equal to .30.

Figure 1 shows three ICCs. The horizontal axis is scaled in units of ability (Θ), and the vertical axis is the probability of answering the item correctly. The solid curved line shows an ICC for an item of average difficulty with acceptable discrimination and the lower asymptote appropriate for a five-item multiple-choice item. The dashed line shows an item of identical difficulty, c value of .28, but with a lower a value. Note how the slope of the curve is less steep. The third curve, dot-dash line, shows an item with a c value of .30, an a parameter of 1.0, and the b parameter equal to 1.0. As the b parameter changes, the location of the inflection point of the curve is displaced along the horizontal axis.

In most cases the test constructor is faced with the task of estimating three parameters for the n items and one ability parameter (Θ) for every examinee (N) so that $N + 3n$ parameters must be estimated for each group of test items. For a group of 2,000 examinees taking 80 items, 2,240 [2,000 + (3 x 80)] parameters must be estimated simultaneously. In an iterative procedure, this estimation must be repeated several times which leads to long computer runs with more precise estimates. Three of the four ICC estimation procedures evaluated in this study are iterative. The fourth is a monotonic increasing function of the biserial correlation between the item and raw score.

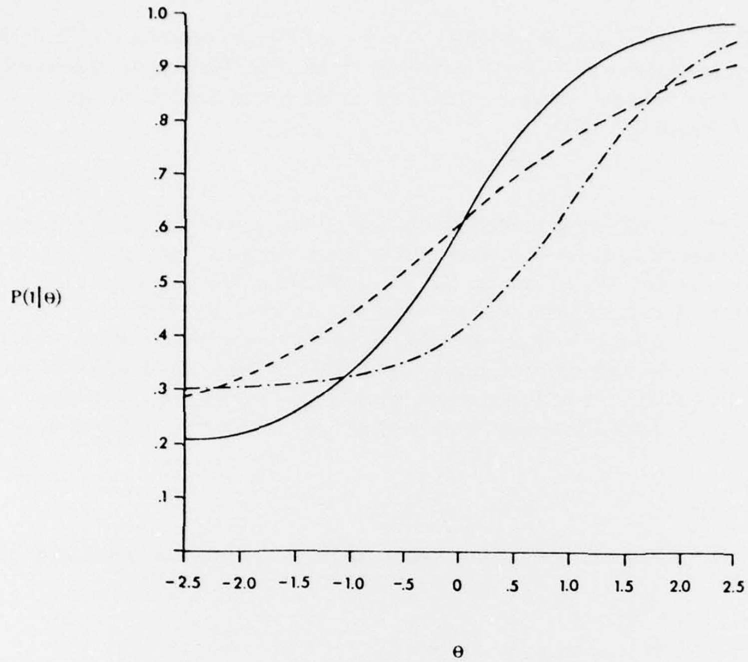


Figure 1. **Item characteristic curves.**

II. METHOD

A simulation was run in order to have known values for ability level (Θ) and for the item parameters. Three distributions of ability (Θ) with differing shapes were generated on which to test the procedures for ICC parameter estimation. Each Θ is equivalent to a "subject." The generated item parameters (a, b, c) remained constant across the three distributions of ability (Θ).

Four methods of assessing the adequacy of the ICC estimation procedures were used. First, the estimated item parameters ($\hat{a}$, $\hat{b}$, $\hat{c}$) were correlated with the known item parameters; second, the Θ estimated by using $\hat{a}$, $\hat{b}$, and $\hat{c}$ from each estimation procedure was correlated with the known Θ. Third "true scores" and estimated "true scores" from the $\hat{a}$, $\hat{b}$, and $\hat{c}$ were compared (Lord, 1975). Finally, the test information curve was compared with estimates of the test information curve using the item parameters estimated in the three data sets. Table 1 shows the means, standard deviations, and minimum and maximum Θ for the three data sets.

Table 1. **Descriptive Statistics for the Distribution of Θ for the Three Data Sets**

Data Set	Mean	Standard Deviation	Minimum	Maximum	Skew	Kurtosis
1	-.0012	1.4437	-2.50000	2.4975	.0000	1.7991
2	.4957	.6998	-.5064	2.3791	.6359	2.7302
3	.0126	1.0191	-3.8445	3.6685	-.0050	3.1144

Data Set 1 (DS1)

The distribution of Θ for *DS1* was generated by dividing the interval between −2.5 and +2.5 into 2,000 equal intervals and assigning each resultant number as a value of Θ. This data set is similar to those sometimes produced for item analytic studies for tests such as the Armed Services Vocational Aptitude Battery (Jensen, Massey, & Valentine, 1976).

Data Set 2 (DS2)

The distribution of Θ for *DS2* was generated by obtaining 3,000 cases from a unit normal random number generator. Two thousand values for Θ were selected by administering a "test" and generating a sum of the number-right scores for the 3,000 based on ICC parameters of a 30-item subtest used in military selection and classification. A cutting score was set which would yield the upper two-thirds of the population. This method, rather than just cutting at a $\Theta \equiv 33.\bar{3}$ percentile equivalent, was used to emulate actual selection practices which involve errors of measurement. The resultant distribution does not have a sharp truncation of Θ but is asymmetric with few scores below a specified level. *DS2* is similar to samples frequently available to organizations which must work with samples selected for inclusion in training or education.

Data Set 3 (DS3)

The distribution of Θ for *DS3* was generated by accessing the unit normal random number generator for 2,000 numbers.

ICC Parameters

The distributions of ICC parameters were generated to simulate 80 five-option multiple-choice test questions. A normal distribution was specified for each ICC parameter. The means and standard deviations of these distributions were set to produce item parameters similar to those likely to be obtained in actual practice. Table 2 describes these distributions.

Table 2. **Descriptive Statistics of the Generated ICC Parameters**

ICC Parameter	Mean	Standard Deviation	Minimum	Maximum
a	.9504	.2837	.4647	1.6136
b	.1635	.9286	−1.6530	1.9745
c	.2009	.0458	.0872	.3479

Note. — These ICC parameters were used for all three data sets.

Generation of Item Responses

In order to generate a vector of item responses for each "subject" the Θ values were used in equation (1) to compute the likelihood of "passing" each item. The three parameter logistic model is given by:

$$P(\Theta)_j = c_i + (1 - c_i)(1 + e^{(-1.7a_i(\Theta - b_i))}) - 1 \qquad (1)$$

where $P(\Theta)_j$ is the probability of "subject" j answering the test item correctly and a_i, b_i, and c_i are item parameters for item i.

Because equation (1) yields a number $P(\Theta)_j$ such that $0.0 < P(\Theta)_j < 1.0$, *a* number, X_j, is drawn from a uniform (rectangular) distribution ranging from 0.0 to 1.0 and compared to $P(\Theta)_j$. If X_j is larger than $P(\Theta)_j$, then an incorrect response is specified for the item; otherwise, a correct response is specified for the item. Thus, a "subject" with $P(\Theta)_j = .90$ gets the item correct 9 in 10 times, and a vector of item responses is developed for each "subject" in each data set. These response vectors are then used to estimate *a*, *b*, and *c* by the four methods.

Estimation of ICC Parameters

The following four methods of ICC estimation were selected because of their wide availability to practitioners: ANCILLES, LOGIST, OGIVIA, and transformations to the item-test biserial correlation. All are three-parameter models.

ANCILLES and OGIVIA (developed by U. S. Civil Service Commission) are described by Urry (1977, 1978) and LOGIST (developed by Educational Testing Service) is described by Wood, Wingersky, and Lord (1976). The transformations may be found in Lord and Novick (1968). These procedures were implemented on a UNIVAC 1108 and thoroughly checked out by processing the sample data set supplied by each of the authors of the programs. Default options for the programs were specified where possible, and the logistic model was used throughout.

III. RESULTS

The first set of analyses consisted of correlating the ICC parameters with the estimated ICC parameters ($\hat{a}, \hat{b}, \hat{c}$). Table 3 shows these results for each data set.

Table 3. **Correlations of ICC and Estimated ICC Parameters**

	ANCILLES			LOGIST			OGIVIA			Transformation		
Data Set	$r_{a.\hat{a}}$	$r_{b.\hat{b}}$	$r_{c.\hat{c}}$	$r_{a.\hat{a}}$	$r_{b.\hat{b}}$	$r_{c.\hat{c}}$	$r_{a.\hat{a}}$	$r_{b.\hat{b}}$	$r_{c.\hat{c}}$	$r_{a.\hat{a}}$	$r_{b.\hat{b}}$	$r_{c.\hat{c}}$
1	.873	.960	.409	.895	.978	.557	.868	.965	.362	.592	.963	*
2**	.440	.941	.027	.565	.447	.233	.556	.923	.000	.323	.917	*
3	.836	.968	.325	.827	.975	.379	.837	.976	.225	.349	.965	*

*Constant value of c = .20 precludes calculation of correlation.
**Entries for ANCILLES and OGIVIA based on 75 and 64 items, respectively.

The second set of analyses was of the correlation of Θ and $\hat{\Theta}$ computed using a maximum likelihood method and the various estimates of a, b, and c from the four procedures. These correlations were analyzed to determine how accurately Θ could be estimated from $\hat{a}, \hat{b}$, and $\hat{c}$ as would be done in adaptive testing.

Maximum Likelihood Estimation (MLE) of Θ is computed using the likelihood function defined as:

$$L(\Theta) = \Pi(P(\Theta)^{u}\, Q(\Theta)^{1-u}) \qquad (2)$$

where $Q(\Theta) = 1 - P(\Theta)$ and u is 1 if the item was answered correctly and 0 if answered otherwise. The maximum of the distribution of likelihoods is found by the method derived by Jensema (1974). The use of this procedure is advantageous because it allows the estimation of Θ regardless of the sequence of item administration. Other methods, such as Bayesian estimation of Θ, are sequence dependent (see Sympson, 1976).

MLE is not sequence dependent but has the problems of possible failure to converge or of reaching an asymptotically infinite estimate. Both of these problems can be rectified by arbitrarily placing a limit on the number of iterations and by placing an upper and lower limit on $\hat{\Theta}$. Maximum Likelihood Estimates of Θ were computed using the response vectors generated from equation (1), each set of estimated item parameters, and the generated item parameters. The estimation of $\hat{\Theta}$ using the generated (a, b, c) item parameters indicates the bias involved in the estimation of Θ alone. The correlation of Θ and the resultant $\hat{\Theta}$ is a measure of test reliability. No correlation of Θ and $\hat{\Theta}$ using any of the estimated $\hat{a}, \hat{b}$, or $\hat{c}$ parameters should be expected to exceed the correlation of Θ and $\hat{\Theta}$ using the generated a, b, c. Table 4 shows the results of these analyses. The column headed Population is the analysis using the generated item parameters.

Table 4. **Descriptive Statistics for the Estimates of Θ Computed from the Generated and Estimated Item Parameters**

(N = 2,000)

	Estimation Method					
	S[a]	Population	ANCILLES	LOGIST	OGIVIA	Transformation
			Rectangular Data Set			
Number of Items	80	80	80	80	80	80
$\bar{X}\hat{\Theta}$	46.257	.0181	.0147	−.0133	.1004	−.0412
$\sigma\hat{\Theta}$	19.629	1.4695	.9223	1.0163	.9087	.9038
$r\Theta.\hat{\Theta}$	.977	.980	.970	.974	.974	.955
$\overline{(\Theta-\hat{\Theta})}$		.0194	.0125	−.0121	.1016	−.0400
μ_{Θ} = −.00125						
Σ_{Θ} = 1.4437						
			Skewed and Selected Data Set			
Number of Items	80	80	75	80	64	80
$\bar{X}\hat{\Theta}$	52.565	.5028	−.0167	.0316	−.4219	.0199
$\sigma\hat{\Theta}$	11.313	.7483	1.0147	1.0263	.9174	.9747
$r\Theta.\hat{\Theta}$	.939	.948	.935	.943	.937	.930
$\overline{(\Theta-\hat{\Theta})}$		−.0071	−.5123	−.4641	−.9176	−.4758
μ_{Θ} = .49574						
Σ_{Θ} = .69989						
			Normal Data Set			
Number of Items	80	80	80	80	80	80
$\bar{X}\hat{\Theta}$	45.587	.0096	.0078	−.0073	.0706	−.0038
$\sigma\hat{\Theta}$	14.615	1.0362	1.0020	1.0147	.9899	1.2313
$r\Theta.\hat{\Theta}$	.957	.966	.964	.965	.965	.961
$\overline{(\Theta-\hat{\Theta})}$		.0223	.0204	.0053	.0833	.0088
μ_{Θ} = −.01269						
Σ_{Θ} = 1.0191						

[a]Indicates number-right score and the all descriptive statistics referred to the number-right score. The correlation is between Θ and S.

The third set of analyses follows guidance proposed by Lord (1975) to eliminate most of the problems associated with estimating extreme values of Θ. These are termed true score (ξ) analyses. Because MLE procedures tend to exhibit bias on extreme cases, there may be a piling-up of high values at the minimum and maximum values allowed by the particular estimation routine. There are no empirical rules for setting either minimum or maximum values to be obtained in the MLE process. The limits set depend on judgment. In this study, the values were set at −2.50 and +2.50. Other values might have yielded slightly different values in Table 4. Estimation of true scores avoids these problems. Equation 3 defines true score.

$$\xi_j = \sum_{i=1}^{n} P_i(\Theta) \qquad (3)$$

where ξ_j is the true score, n is the number of items, and $P_i(\Theta)$ is the probability of a correct response for

the item as in equation (1). Similarly, the estimated true score is given by

$$\hat{\xi}_j = \sum_{i=1}^{n} P_i(\hat{\Theta}) \tag{4}$$

where $P_i(\hat{\Theta})$ is computed from equation (1) using $\hat{a}$, $\hat{b}$, and $\hat{c}$.

Table 5 shows the means and standard deviations of ξ and $\hat{\xi}$, the average difference between them, and their intercorrelation.

Table 5. **Descriptive Statistics of ξ and $\hat{\xi}$, the Average Difference Between Them and Their Correlation**

Procedure	$r_{\xi.\hat{\xi}}$	$\bar{X}_{\hat{\xi}}$	$s_{\hat{\xi}}$	$\overline{(\xi - \hat{\xi})}$
		Data Set 1		
ANCILLES	.9927	46.444	24.927	.3444
LOGIST	.9960	47.205	23.424	1.1059
OGIVIA	.9945	45.210	25.091	−.8895
Transformation	.9910	47.589	24.352	1.4894
μ_ξ = 46.099				
σ_ξ = 19.245				
		Data Set 2		
ANCILLES[a]	.9995	54.63	7.783	5.3617
LOGIST	.9997	58.02	7.260	5.531
OGIVIA[b]	.9994	45.52	7.7895	−.4415
Transformation	.9999	58.04	8.028	5.550
μ_ξ = 52.49				
σ_ξ = 10.592				
		Data Set 3		
ANCILLES	.9998	45.90	14.325	.5737
LOGIST	.9999	46.085	14.112	.7591
OGIVIA	.9999	45.158	14.044	−.1680
Transformation	.9999	45.950	14.157	.6236
μ_ξ = 45.326				
σ_ξ = 14.204				

[a]75 items only for ξ and $\hat{\xi}$.

[b]64 items only for ξ ans $\hat{\xi}$.

The fourth set of analyses consisted of comparisons of the test information curve using the known *a b*, *c* versus test information computed from $\hat{a}$, $\hat{b}$, $\hat{c}$ from the four-item parameter estimation techniques.

Item information is defined as

$$I_g(\Theta) = \frac{(\frac{\partial}{\partial\Theta} P_g(\Theta))^2}{P_g(\Theta)(1 - P_g(\Theta))} \tag{5}$$

where $P_g(\Theta)$ is estimated from equation (1) and the numerator is the squared first derivative (i.e., the squared slope) of $P_g(\Theta)$ at a fixed value of Θ. Test information is the sum of the item information curves making up a test and is defined as

$$I(\Theta) = \sum_{i=1}^{n} I_g(\Theta) \tag{6}$$

where $I_g(\Theta)$ is defined in equation (5). Estimates of item information ($\hat{I}$) may be computed by substituting $\hat{a}, \hat{b}, \hat{c}$ into equation (1) and substituting that quantity into equations (5) and (6).

It is useful to calculate item and test information curves in order to determine the precision of measurement of a test or an item. The height of the item or test information curve at any level of Θ may be thought of as being an ICC analog to classical measures of reliability. The higher the information curve the higher the information value and the higher the reliability of the item or test at that level of Θ.

Test information curves are frequently used to compare test characteristics (Brown & Weiss, 1977; McBride & Weiss, 1976; Vale & Weiss, 1977; and Weiss, 1975) and to select items for administration during adaptive testing (Jensema, 1974; Ree, 1977). Because test and item information curves are computed using ICC parameters, errors of estimation of the parameters can cause errors in the test and item information curves.

Figures 2, 3, and 4 show the test information curve and estimates of the test information curve based on $\hat{a}, \hat{b}, \hat{c}$ estimated by the four methods in each of the data sets. The item parameters have been made comparable by placing them on common metric via a linear transformation of a and b. No such transformation of c is necessary. Table 6 presents the sum of squared deviations of true test information minus estimated test information as well as the point on Θ where information reaches its maximum (Θ_g), the correlation of I and $\hat{I}$, and minimum and maximum values of $\hat{I}$ computed by each method in each of the data sets.

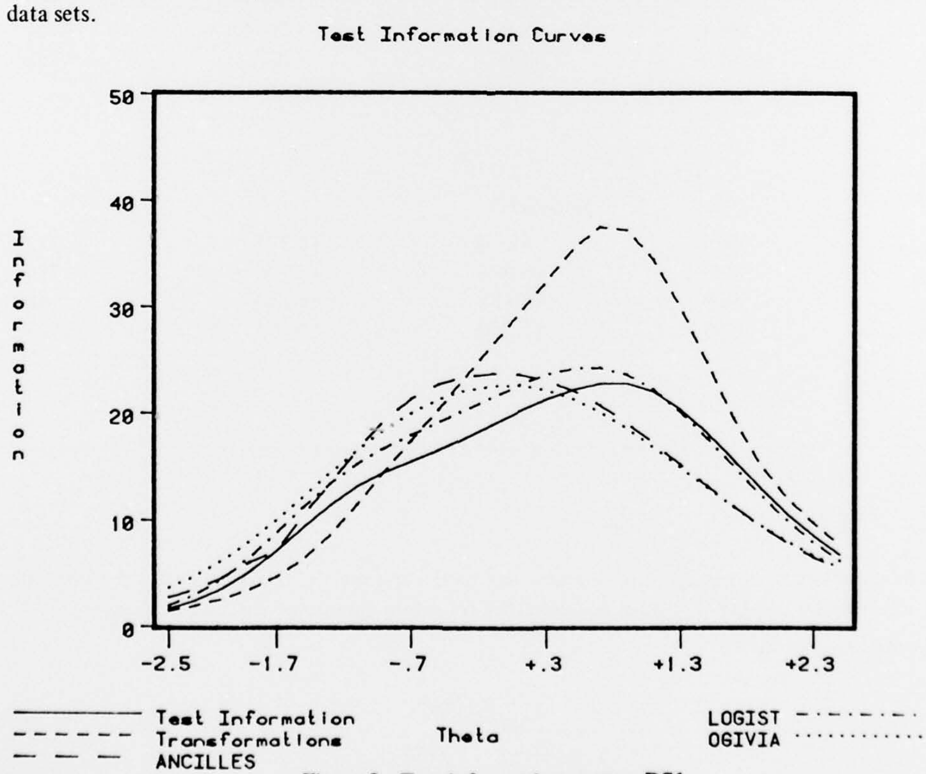


Figure 2. **Test information curves, DS1.**

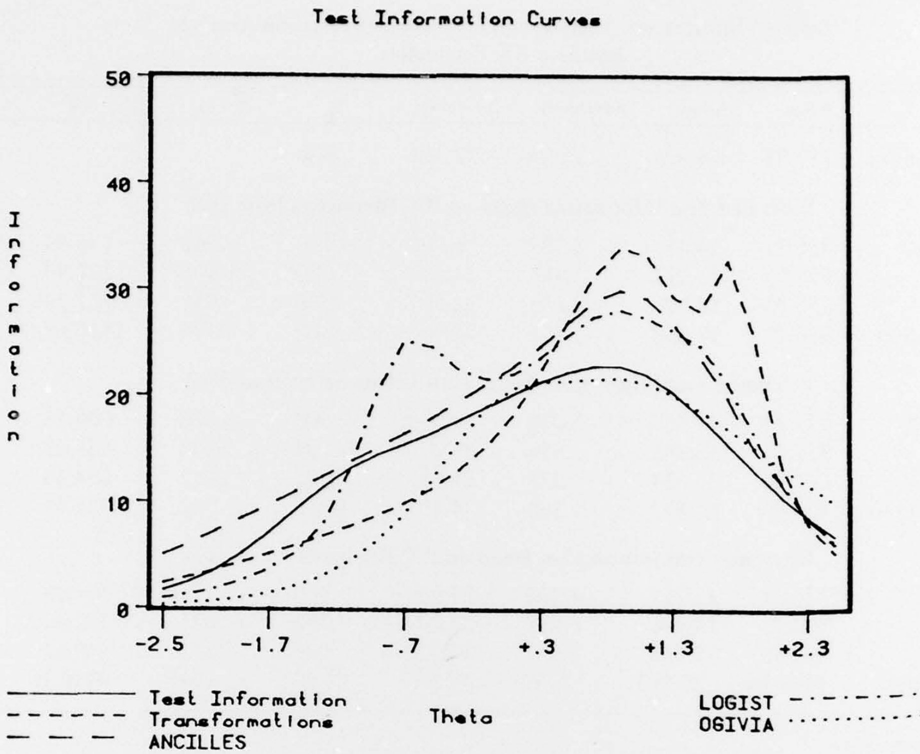


Figure 3. **Test information curves, DS2.**

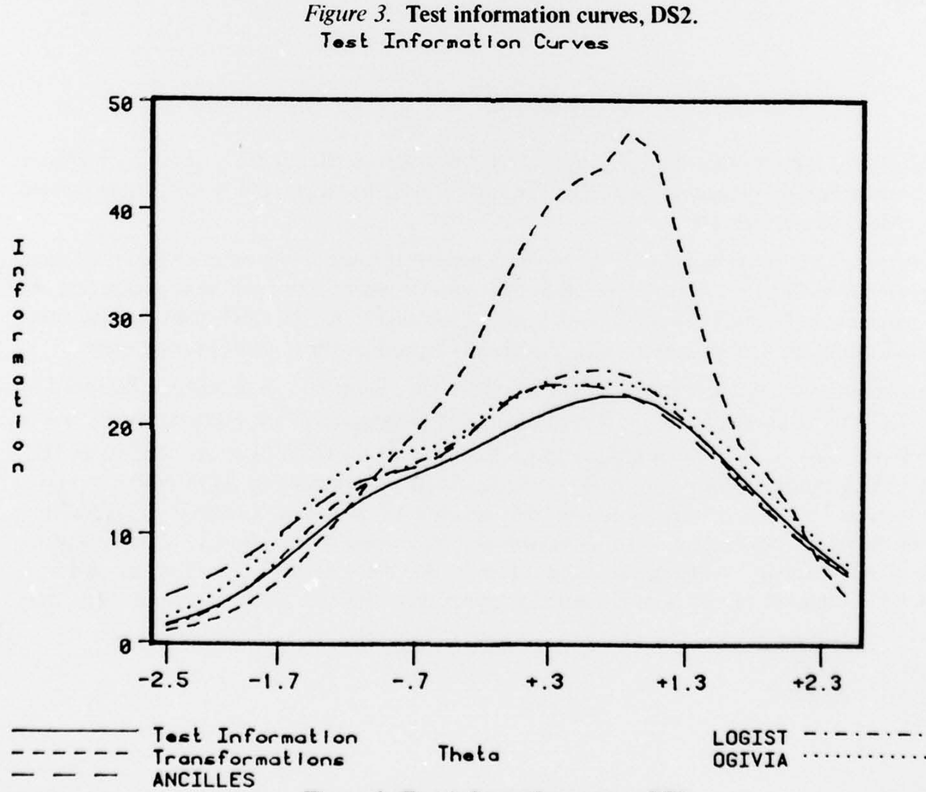


Figure 4. **Test information curves, DS3.**

Table 6. **Information Analysis and Estimated Information Analyses Based on ICC Parameters**

	Total	Mean	Minimum	Maximum	Θ_g	$\overline{(\hat{I} - I)}$	$\Sigma(I - \hat{I})^2$	$r_{I.\hat{I}}$
Test Inforamtion	717.83	14.075	1.695	22.924	.800			
Estimated Test Information Based on ICC Parameters from DS1								
ANCILLES	736.31	14.437	2.757	23.708	–.100	–.362	650.04	.864
LOGIST	775.76	15.211	1.989	24.314	.600	–1.136	137.96	.986
OGIVIA	735.75	14.426	3.658	22.607	.000	–.351	621.29	.850
Transformation	930.77	18.250	1.477	37.708	.800	–4.175	2510.61	.971
Estimated Test Information Based on ICC Parameters from DS2								
ANCILLES[a]	871.24	17.083	5.016	29.954	.900	–3.008	694.15	.970
LOGIST	835.54	16.383	.954	28.096	–.600	–2.308	989.93	.958
OGIVIA[b]	613.24	12.024	.338	21.682	.900	2.051	854.63	.899
Transformation	806.66	15.817	2.360	34.105	1.00	–1.742	2361.61	.821
Estimated Test Information Based on ICC Parameters from DS3								
ANCILLES	777.81	15.251	4.280	23.906	.400	–1.1760	174.48	.976
LOGIST	762.35	14.948	1.539	25.300	.700	–.873	102.67	.994
OGIVIA	812.85	15.938	2.332	24.022	.800	–1.863	219.61	.991
Transformation	1070.70	20.993	1.046	47.565	1.00	–6.918	6416.10	961

[a] 75 items only.

[b] 64 items only.

IV. DISCUSSION

The results clearly indicate that no one program functions best in all situations posed by the three data sets. The transformation procedure performed poorly in most instances and is not recommended unless no other procedures are available.

In the rectangular data set (*DS1*), LOGIST produces results superior to the other procedures except in terms of the average differences between ξ and $\hat{\xi}$. The correlations of estimated item parameters and generated item parameters, Θ and $\hat{\Theta}$, I and $\hat{I}$, and ξ and $\hat{\xi}$, are higher for LOGIST than for any other procedure. LOGIST estimated item parameters also most nearly reproduce the test information curve.

The results from the skewed and selected data set, *DS2*, call attention to a peculiarity exhibited by ANCILLES and OGIVIA. Under specific conditions, these two programs will not estimate parameters of some items. While this may seem a disadvantage, notice that $\overline{(\xi - \hat{\xi})}$ for OGIVIA is the smallest in *DS2*. Note also that OGIVIA shows (Table 4) an $r\Theta.\hat{\Theta}$ of .937 for 64 items compared to .943 for 80 items using LOGIST. This increase of .006 is very small for the addition of 16 items. LOGIST estimates item parameters for all the items, but inspection of the scatter plot of b versus $\hat{b}$ indicates several outliers which have the effect of substantially reducing the value of $rb.\hat{b}$. All the estimated test information curves computed from *DS2* estimates of the item parameters approximate the true test information curve very poorly.

The OGIVIA procedure is the most preferable for use in the normally distributed data set, *DS3*. The correlations of OGIVIA estimated $\hat{a}$ and $\hat{b}$ with a and b are higher than for the other procedures; however, its correlation of c and $\hat{c}$ is less than that of either ANCILLES or LOGIST. The $r\Theta.\hat{\Theta}$ using OGIVIA is as high as LOGIST and higher than all others. The $r\xi.\hat{\xi}$ for OGIVIA is the highest and simultaneously has the smallest average difference between ξ and $\hat{\xi}$. OGIVIA is built around assumptions of the normality of the distribution of Θ and performs very well when these conditions hold true, as in *DS3*, or approximately hold true, as in *DS2*. LOGIST estimates of the item parameters produce the highest correlation between I and $\hat{I}$ and the lowest sum of squared deviants of I minus $\hat{I}$ and thus the best estimate test information.

The decision as to which procedure to use must be based on a series of criteria. If all the items must be calibrated, then OGIVIA and ANCILLES may present problems in a situation like that represented by *DS2*. If wide range samples like *DS1* and *DS3* are available and the estimation of Θ is the goal, then calibration with LOGIST or OGIVIA is recommended. Clearly, if the examinees are available, a normal distribution of Θ leads to the best estimations of a, b, c, ξ, Θ, I and is desirable. These data should then be calibrated using OGIVIA.

A final factor should be considered: cost. The transformation procedure was the quickest because, unlike the others, it is not iterative and its work can be accomplished in about 10 FORTRAN statements. The LOGIST procedure takes the longest on the computer. It ran eight times longer than either ANCILLES or OGIVIA. Central Processor Unit (CPU) times on a UNIVAC 1108 with 262K words of memory for *DS3* were for ANCILLES, 296 seconds; LOGIST, 2,061 seconds; OGIVIA, 180 seconds; and transformations, 38 seconds.

The choice of ICC parameter estimation techniques should be consistent with the planned use of the estimates, the characteristics of the distribution of ability in the groups available for item administration, the necessity to calibrate all items, and the computer resources available.

REFERENCES

Brown, J.M., & Weiss, D. *An adaptive testing strategy for achievement test batteries.* Research Report 77-6. Minneapolis, MN: University of Minnesota, 1977.

Jensema, C. An application of latent-trait mental test theory. *British Journal of Mathematical and Statistical Psychology*, 1974, **27**, 29–48.

Jensen, H., Massey, I., & Valentine, L. *Armed Services Vocational Aptitude Battery Development (ASVAB Forms 5, 6, and 7).* AFHRL-TR-76-87, AD-A037 522. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, December 1976.

Lord, F. *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters.* Research Memorandum 75-33. Princeton, NJ: Educational Testing Service, 1975.

Lord, F., & Novick, M. *Statistical theories of mental test scores.* Reading, MA: Maddison-Wesley, 1968.

McBride, J., & Weiss, D. *Some properties of a Bayesian adaptive ability testing strategy.* Research Report 76-1. Minneapolis, MN: University of Minnesota, 1976.

Ree, M. Implementation of a model adaptive testing system at an Armed Forces Entrance and Examination Station. *Proceedings of the 1977 Computerized Adaptive Testing Conference*, Minneapolis, Minnesota, July 1977.

Symposon, J. Estimation of latent trait status in adaptive testing procedures. *Proceedings of the 18th Annual Convention of the Military Testing Association*, Gulf Shores, Alabama, 1976.

Urry, V. *OGIVIA: Item parameter estimation program with normal ogive and logistic three-parameter model options.* U. S. Civil Service Commission, Washington, D.C., 1977.

Urry, V. *ANCILLES: Item parameter estimation program with normal ogive and logistic three-parameter model options.* U. S. Civil Service Commission, Washington, D.C., 1978.

Vale, C., & Weiss, D. *A comparison of information functions of multiple-choice and free-response vocabulary items.* Research Report 77-2. Minneapolis, MN: University of Minnesota, 1977.

Weiss, D. Adaptive testing research at Minnesota: Overview, recent results, and future directions. In *Proceedings of the First Conference on Computerized Adaptive Testing.* Washington, D.C.: U. S. Government Printing Office, 1975.

Wood, R., Wingersky, M., & Lord, F. *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters.* Research Memorandum 76-6. Princeton, NJ: Educational Testing Service, 1976.

☆U.S. GOVERNMENT PRINTING OFFICE:1979- 671-056/103